LLNL-TR-560431

LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

Determining mutant spectra of three RNA viral samples using ultra-deep sequencing

H. Chen

June 8, 2012

Disclaimer

This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

BMI 217 Final Project Haiyin Chen

Determining mutant spectra of three RNA viral samples using ultra-deep sequencing

Abstract

RNA viruses have extremely high mutation rates that enable the virus to adapt to new host environments and even jump from one species to another. As part of a viral transmission study, three viral samples collected from naturally infected animals were sequenced using Illumina paired-end technology at ultra-deep coverage. In order to determine the mutant spectra within the viral quasispecies, it is critical to understand the sequencing error rates and control for false positive calls of viral variants (point mutantations). I will estimate the sequencing error rate from two control sequences and characterize the mutant spectra in the natural samples with this error rate.

Acknowledgement

This work is part of a DTRA-funded viral evolution study headed by Monica Borucki at Lawrence Livermore National Lab. I worked closely with Jonathan Allen on bioinformatics aspects of the study. Jonathan Allen generated the nucleotide frequency distribution (.freq) files and the quality score distribution (.qscores) files, which served as the starting point of my analysis for this paper. I used two additional programs he developed in my analysis (see readme.txt).

Introduction

Viruses with RNA genomes replicate with extremely high mutation rates because their RNA polymerases lack the proofreading ability of DNA polymerases (1). At about 1 error per 10000 nucleotides copied, a point mutation is introduced nearly every time a single RNA virus replicates. Any given host-derived viral sample contains a diverse population of viral strains that are evolutionarily related through mutation. Such cloud of related genotypes is referred to as a quasispeices or a mutant swarm, whose genetic diversity and mutational speed confers its ability to adapt under selection pressure as a whole. Understanding the evolution dynamics of RNA virus is key to our understanding of viral disease progression, transmission and developing antiviral therapeutics.

High throughput sequencing is beginning to make transformative impact on the area of viral evolution. Until recently, studies of viral genomes have mostly focused on the consensus sequence that identifies the predominant viral sequence but is uninformative about the population diversity present in the sample (2). To detect any minority variants, traditional cloning approach had been used where viral sample is diluted and cloned for Sanger sequencing. Not only is this procedure laborious and costly, it also has very limited resolution to detect the real heterogeneity of viral population in the samples. Ultra-deep sequencing enabled by next-generation sequencing platforms has the potential to reveal the mutant spectrum within a sample in high-resolution through massive coverage of the viral genome. The challenge however, is to accurately reconstruct the viral population from the sequencing data and differentiate real viral mutations (from the predominant sequence) in the minority variants from sequencing errors.

As a pilot study for a series of viral evolution studies conducted at Lawrence Livermore National Lab, three viral samples collected from naturally infected hosts --two fox rabies brain tissue samples collected during a 2009 Humboldt County outbreak and one bovine coronavirus nasal sample collected from a calf at a Northern California farm – along with two 1kb plasmid clones were sequenced using Illumina Paired-end Technology at ultra-deep coverage. The plasmid clones serve as sequencing control in the study -- their known sequences provide the ground truth to their sequenced data and help define the combined PCR and sequencing error rate for the natural samples.

To control for sequencing errors and minimize false positive variant calls, we explored several approaches and constrains. First, to take advantage of the Paired-end Technology, only overlapping regions of the read pairs were used to make variant SNP calls in this paper. The overlapping read pairs provide an added layer of error-checking. Second, we examined the relationship between quality scores and mismatch errors in the read pairs and developed ways to select quality score cutoff. Third, we incorporated mismatch error rates in the read pairs when making variant SNP calls (i.e. point mutations) so that the modified hypothesis testing is sensitive to local error rate.

One first step to understanding viral mutational dynamics involves determining whether point mutations evolved in isolation or in conjunction with each other. Without carrying out haplotype reconstruction (3), the range we can make inferences between point mutations is limited by the Illumina sequencing read lengths. We can, however, test genotype association wherever variant SNPs are covered by the same reads. In this paper we test genotype association between pairs of two nearby viral variants discovered by our algorithm.

Method

Data

Five RNA viral samples were sequenced using Illumina Paired-end Technology for this study. Each sample took up 1 lane of an Illumina flowcell. The two control samples are: a 1kb plasmid clone containing a fragment of BCV virus and a 1kb plasmid clone containing a fragment of rabies virus. The three natural samples are one bovine coronavirus (BCV) sample prepared from nasal swab of a calf and two rabies samples prepared from brain tissues of two foxes. All samples were PCR amplified before sequencing. Only overlapping regions in the read pairs were considered for the present study. Analyses on the natural samples were carried out in sample-coordinates.

Determining quality score cutoff

At every base, all overlapping read pairs were separated into two categories: matching and non-matching pairs. Matching pairs have two complementary nucleotides and non-matching pairs have two incongruent nucleotides. Quality scores (Q) for every base pair were compiled as follows. For matching read pairs, the average quality score was used. For non-matching pairs, the minimum quality score was used. Resulting two Q-score distributions were compared and used to generate Q-score receiver-operator characteristic (ROC) and 'false discovery rate' curves.

Sequencing errors can occur in two forms in overlapping read pairs. Non-complementarity between the forward and reverse strands at a given base indicates that at least one of the two nucleotides is erroneously incorporated. This type of error is easy to exclude, as the non-matching read pairs are excluded from analysis except those for quality control. A second, more rare but 'hidden' form of error is where two complementary errors occur on both the forward and the reverse strands such that the resulting read pair appears perfectly matched.

Making variant SNP calls

We make variant SNP calls using hypothesis test based on the binomial distribution, where the probability of observing x or more mutations in N matching read pairs covering the base is given by the survival function of the binomial distribution $B(N, p)$

$$\mathrm{P}(X \geq x) = \sum_{k=x}^{N} \binom{N}{k} p^k (1-p)^{N-k}$$

The probability of sequencing error, p, is the combined PCR and sequencing error, ε, adjusted by a function of the read-pair mismatch rate, δ, at the base in question.

The read-pair mismatch rate, δ, is the *position-dependent* rate at which a nucleotide is mis-incorporated into a single strand. For simplicity, we modeled the probability at which two complementary nucleotides are mis-incorporated simultaneously on forward and reverse strands at the same base as δ^2. Hence, we used the maximum of clonal control derived error and sequence-specific position-dependent mismatch rate as the adjusted *base-dependent error rate* for the paired end reads.

$$p = max(\varepsilon, \delta^2)$$

P-value = 0.01 with Bonferroni correction was used as the significance threshold. Q-value method by Storey (4) turned out to be not conservative enough.

Linkage analysis of adjacent variant SNP pairs

Variant SNPs were ordered according to their locations on the genome. Adjacent, or neighboring variant SNPs could be 1 bp or hundreds of bp apart. For every pair of adjacent variant SNPs, distribution of all genotypes covering the two loci were gathered from all reads spanning the two loci. For two given loci, a contingency table was constructed as below where C1, C2 were the consensus nucleotides at locus 1 and 2, respectively, and V1, V2 were the variant SNP nucleotides at locus 1 and 2, respectively. The number of reads carrying each of the four genotypes of interest (other genotypes are possibly present since a viral quasispecies is multiploid) was entered into the table.

# reads with genotype	**C2**	**V2**
C1	C1--C2	C1--V2
V1	V1--C2	**V1--V2**

Fisher's exact test was used to evaluate the statistical significance of these contingency tables, where the null hypothesis is that the two classifications, C1/V1 and C2/V2, are not associated. If the null hypothesis is rejected and the proportion of reads covering V1-- V2 is large in comparison to those covering C1--V2 and C2--V1, then it is likely that V1 and V2 are linked.

Results

Figure 1 shows the coverage levels for the two control samples in terms of raw read count (blue). The raw reads at a given base are unfiltered by quality scores. Some are in the overlapping regions of read pairs, others are in the non-overlapping regions of read pairs. Some read pairs match, others do not. Figure 1 also shows coverage level by matching read pairs thresholded at Q ≥30 (in units of pairs). This filtering removes a large fraction of the raw reads. The resulting average coverage levels by the matching read pairs for the BCV and rabies plasmid controls are 116,010x and 581,855x read pairs, or 232,020x and 1,163,710x single reads. (All coverage information is given as read pairs from here on.)

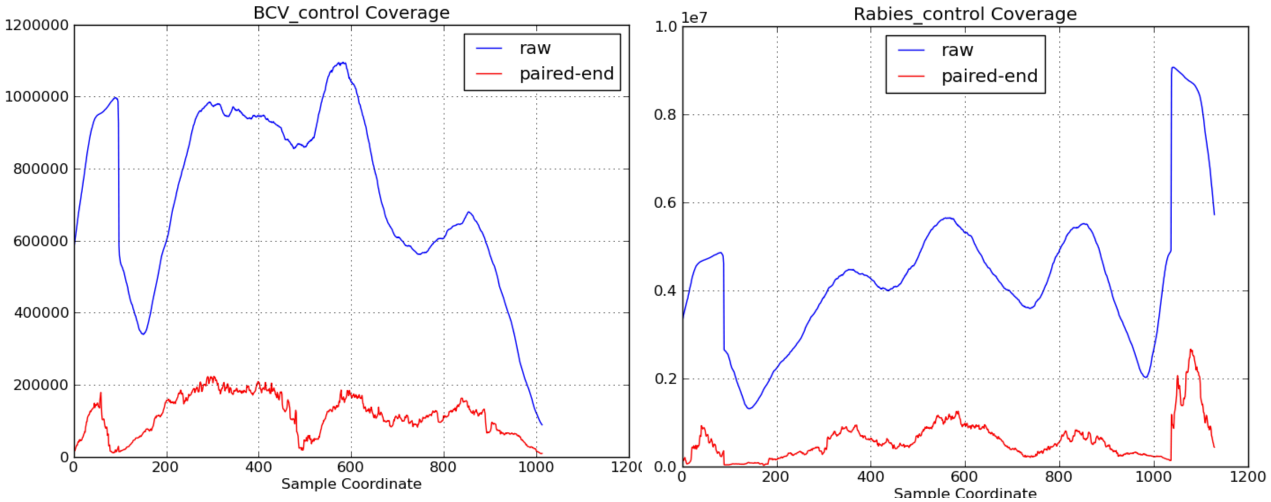

Figure 1 Coverage levels for the two control sequences. Blue: raw reads. Red: quality score filtered paired-end reads, qscore >=30. Since only overlapping regions of the read pairs were used, the actual coverage for the paired-end reads is twice of what is shown in red. Only matching read pairs were used.

Error rates in the two control sequences

Error rates in the two control sequences served as the baseline error rate for the three natural samples. The error rate at a given base in a control sequence is the fraction of reads that do not agree with the known nucleotide at the base, i.e., the fraction of non-consensus reads at the base. The distributions of error rates in the two controls are shown in Figure 2 and Table 1. This error rate is the combined PCR and sequencing error. While the mean error rate is below 2.5e-05, the maximum can be as high as 0.0006.

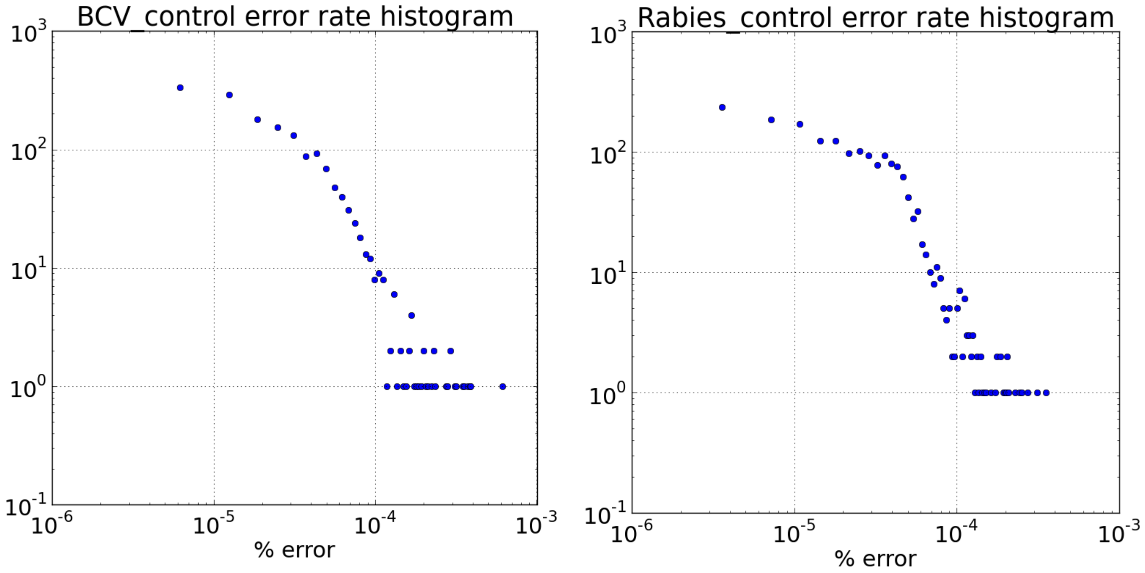

Figure 2 Error rate histograms for the two control sequences

Table 1 Summary of error rates in the two control sequences

Control sample	Mean error rate	Median error rate	Maximum error rate
BCV	2.242e-05	1.119e-05	6.214e-04
Rabies	2.484e-05	1.657e-05	3.597e-04

Q-score analysis

During Illumina sequencing, every base call is issued a Phred-like quality score that indicates the confidence of the call – the higher the quality score, the more confident the base call is correct. The quality score of a given base, Q, is defined by the equation, Q = -10log$_{10}$(e), where e is the estimated probability of the base call being wrong. A Q-score of 30 corresponds to an error rate of 1 in 1000. A Q-score of 35 corresponds to an error rate of 3 in 10,000.

We sought to establish a quality threshold as a read inclusion criteria. To the first order, the matching overlapping read pairs represent correct calls and and mismatched read pairs represent incorrect calls. By comparing Q-scores of these two populations of reads, we can assess what cutoff value might best separate the correct from the incorrect calls. Figure 3 shows in linear scale the Q-score distributions of matching vs. non-matching read pairs from BCV control data. Not surprisingly, the matching read pairs tended to have high Q-scores and nonmatching long Q-scores. However, the reverse is also true for a fraction of read pairs. The fact these two histograms overlap completely means there is no Q-score that can safely separate the 'correct' from the 'incorrect' reads (first order). Low scoring matching pairs are likely candidates for 'hidden' or complementary errors. High scoring mismatching reads suggest even matching reads with high Q-scores will contain complementary errors at some low probability. Although in practice we exclude all non-matching read pairs, nonetheless, they provide a best estimate of the error rate on single strand reads at each Q-score and upper bound for complementary errors in matching reads.

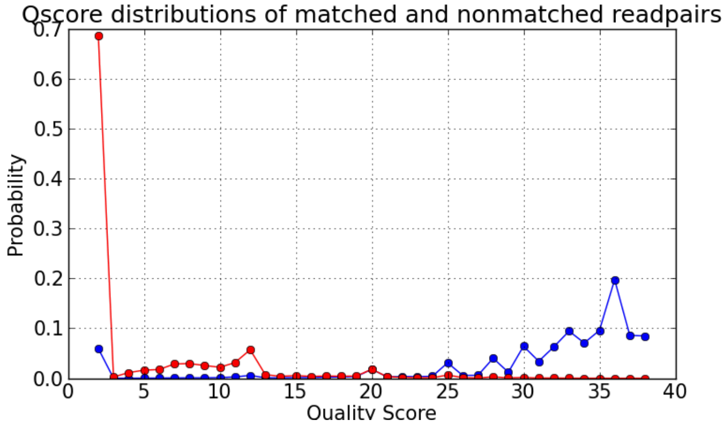

Figure 3 Q-score distribution of matching vs. non-matching reads from BCV control, shown in linear scale. Blue: matching read pairs. Red: non-matching read pairs.

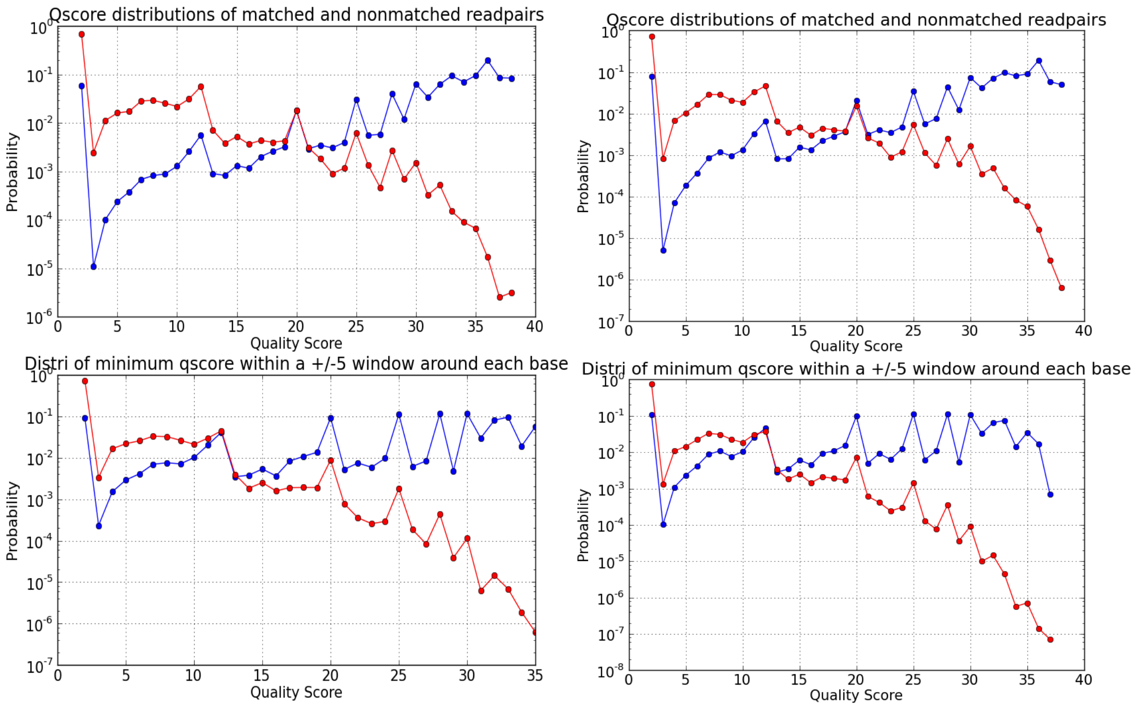

Figure 4 Q-score distributions for matching vs. non-matching read pairs. Left: BCV control data. Right: rabies control data. Top: Q-score as reported at the base. Bottom: minimum Q-score within an 11bp window centered at the base. Blue: matching read pairs. Red: non-matching read pairs.

Figure 4 shows in log-scale the Q-score distributions for matching vs. non-matching read pairs. The two histograms intersect below Q = 25. To be more conservative, instead of using the Q-score provided for each base, we assign the minimum Q-score within bp window to each base (this was applied for subsequent analyses and discussions). In this case, the two histograms intersect below Q = 15. Again, if we view the matching read pairs as proxy for 'positives' and non-matching pairs as 'negatives', then ROC curves are best for summarizing performance of a binary classifier system. Figure 5 shows ROC curves for the two control data sets. At Q = 30 (red dots on black lines), more than 60% of the matching reads are excluded (false negatives) at a false positive rate of 10^{-4}. This FPR of 10^{-4} is our estimate of single read error rate among reads with Q 30, and upper bound for erroneous matching read pairs.

Another way to visualize this information is to plot the 'false discovery rate' at each Q-score q, where FDR is defined as the fraction of reads with Q q (positives) that are mismatched read pairs. This FDR is an upper bound estimated of the fraction of matching reads containing errors, or sequencing error among matching read pairs. At Q=30, FDR is below $3x10^{-6}$. Since this value is below the general estimate of PCR errors, we believe Q=30 is a reasonable threshold for matching read pair s inclusion.

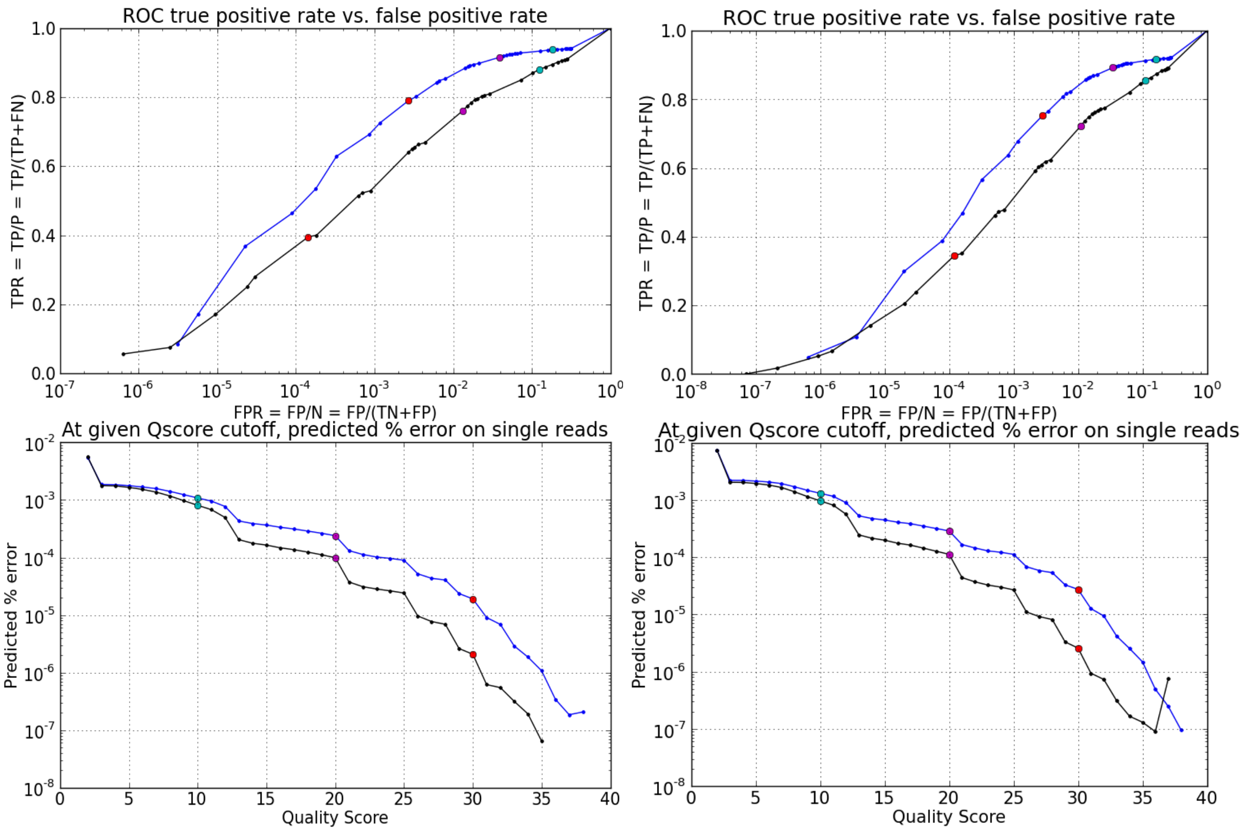

Figure 5 ROC (top) and FDR (bottom) plots based on varying quality score thresholds. Left: BCV control data. Right: Rabies control data. Blue: based on Q-scores at each base. Black: based on minimum Q-score within a 11bp window centered at each base. Red dots: Q-score = 30. Magenta dots: Q-score = 20. Cyan dots: Q-score = 10.

Determining error rate for making variant calls

We applied the variant call algorithm to the control data and found that the error rates estimated earlier using average error rate in the control data (Table 1) and FDR were too low. Table 2 shows the number of SNPs found in the control sequence for various error rates. Based on these results, a conservative error rate of 0.0005 was chosen for making variant SNP calls on the 3 natural samples. We speculate this combined error rate is dominated by the PCR error rate.

Table 2 Number of "SNPs" found in the two control sequences given different error rates

Control Sample	Err rate 0.00005	Err rate 0.0001	Err rate 0.0004	Err rate 0.0005
BCV	30	12	0	0
Rabies	91	27	1	1

Making variant SNPs calls

Total of 160, 107 and 103 variants were called in the BCV, Fox1 and Fox2 samples, respectively (Table 3).

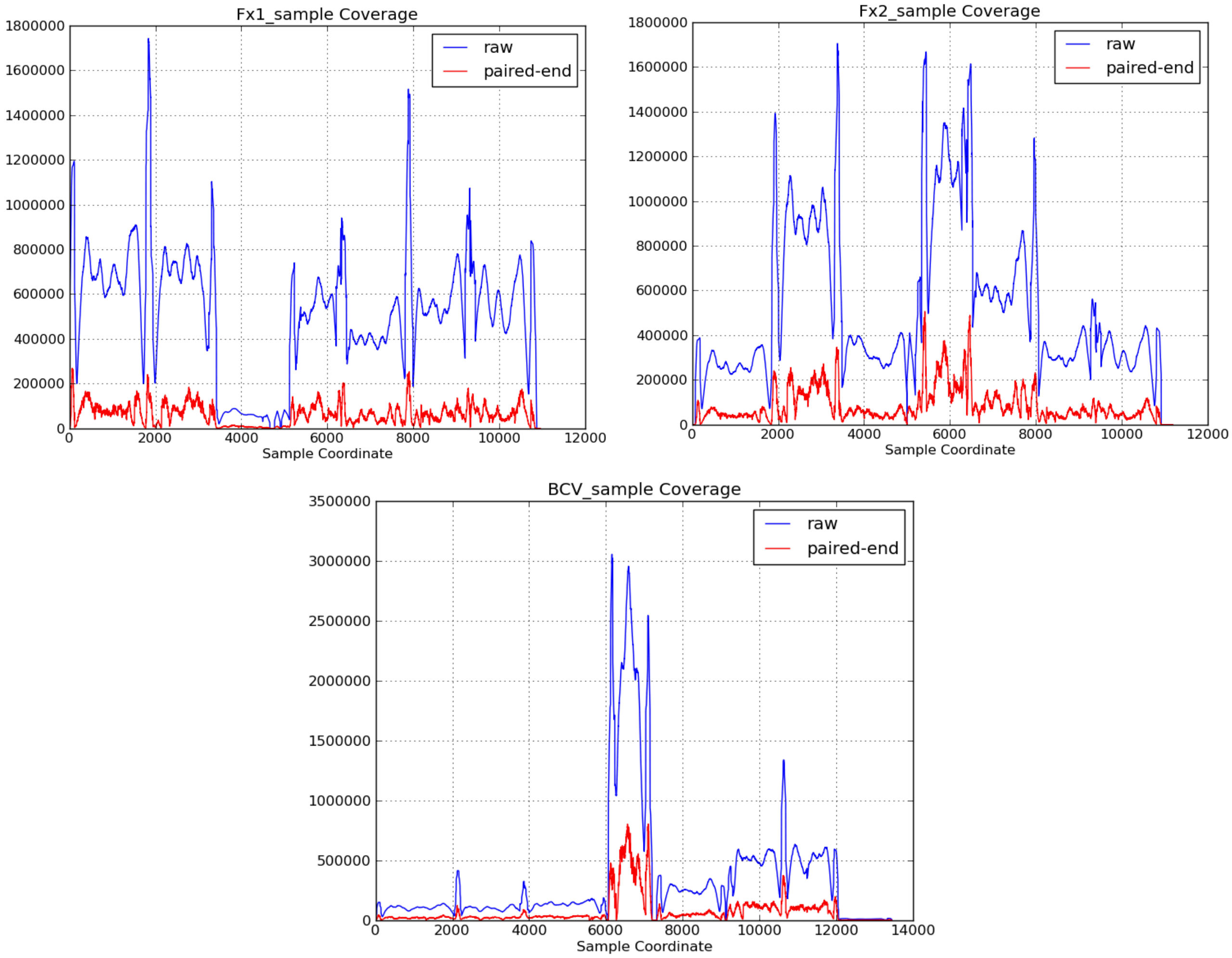

Figure 6 Raw (blue) and filtered paired-end reads coverage (red) for Fox1, Fox2 and BCV samples. Only matching read pairs were considered. Quality score filter for paired-end reads was Q>=30. Coverage for paired-end reads is shown in pairs, i.e., the actual number of reads is twice as many.

Table 3 Summary of coverage, variants and SNPs detected in paired-end reads of 3 natural samples

Natural Sample	# bases sequenced	Mean coverage	# variants in reads filtered at Q=30	# SNPs called at error rate = 0.0005	FDR
BCV	13434	73725x	21284	160	99.25%
Fox1	10905	62171x	20829	107	99.48%
Fox2	11183	85514x	22978	103	99.55%

Variant SNPs linkage analysis

For every pair of neighboring variant SNP loci, Fisher's exact test was used to determine the statistical significance of the genotype distributions involving two variant SNPs and the corresponding consensus nucleotides at the two loci. For example, the 160 SNPs discovered in the BCV sample, form 159 neighboring loci pairs. Table 4 summarizes the significant test results at different p-value thresholds.

Fisher's exact test p-values	# variant pairs in BCV	# variant pairs in Fox1	# variant pairs in Fox2
$10^{-10} < p < 10^{-3}$	2	0	0
$10^{-50} < p < 10^{-10}$	4	0	0
$10^{-100} < p < 10^{-50}$	4	0	0
$p < 10^{-100}$	6	0	0

Linkage between 6 of the 159 pairs of BCV variant SNPs were highly significant. The fox rabies samples however did not have any significantly linked variant SNPs.

Discussion

In general it is difficult to separate PCR and sequencing errors.

References

1. Lauring AS, Andino R (2010) Quasispecies Theory and the Behavior of RNA Viruses. PLoS Pathog 6(7): e1001005. doi:10.1371/journal.ppat.1001005

2. Wright CF, Morelli MJ, Thébaud G, Knowles NJ, Herzyk P, Paton DJ, Haydon DT, King DP. Beyond the consensus: dissecting within-host viral population diversity of foot-and-mouth disease virus by using next-generation genome sequencing. J Virol. 2011 Mar; 85(5):2266-75.

3. Eriksson N, Pachter L, Mitsuya Y, Rhee S, Wang C, Gharizadeh B, Ronaghi M, Shafer RW, & Beerenwinkel N, (2008) Viral Population Estimation Using Pyrosequencing. PLoS Comput Biol, Vol. 4, pp. e1000074

4. Storey JD. (2002) A direct approach to false discovery rates. Journal of the Royal Statistical Society, Series B, 64: 479-498.

Readme.txt

```
--------------------------------------------------------------------------------
Code written by Haiyin Chen
--------------------------------------------------------------------------------
plot_coverage_sample.py    -- plot coverage
errorprofile.py            -- plot histogram of error rates
qscoreHist.py              -- make q-score ROC plot and estimated % error reads
                           for given qscores
call_snps_hc.py            -- compute p-value for variant SNPs based on binomial
                           distribution, generate *.sig files
getNumSigVar.R             -- get number of significant variants in the *.sig
                           file.  p<0.01/Bonferroni correction
BMIproject.R               -- Bonferroni correction of variant SNP p-values,
                           Fisher exact test on variant pair linkage
--------------------------------------------------------------------------------
Code written by Jonathan Allen (LLNL colleague)
--------------------------------------------------------------------------------
freq2error.pl              -- obtain error rates/subconsensus variant frequency
                           in control sequence
rbam_pe_hap                -- generate genotypes for given locus pairs
```